DOCUMENT RESUME

ED 245 703 IR 050 775

AUTHOR Bond, David K.
TITLE Development and Implementation of a Hierarchical Classification System for the ABI/INFORM Database.
PUB DATE May 83
NOTE 10p.; Paper presented at the Mid-Year Meeting of the American Society for Information Science (12th, Lexington, KY, May 22-25, 1983).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administration; Business; *Classification; *Databases; *Information Retrieval; *Online Systems; Periodicals; *Search Strategies; Systems Development; Technological Advancement; Vertical Organization
IDENTIFIERS *ABI INFORM

ABSTRACT

A new search feature has been developed and implemented to enhance online information retrieval from the ABI/INFORM database, which includes abstracts from more than 500 journals covering all aspects of business and management. Published by Data Courier Inc., it includes material dating back to 1971. Because of the diverse nature of the database and the imprecision of management terminology, it was determined that online searchers would be well-served by a hierarchical classification system. Such a system has been developed and implemented over the past two years. Because of the costs involved, it is common in the online industry for such enhancements to be searchable only from the date of implementation. The management of Data Courier Inc., however, believed this feature should operate throughout the database. Classification codes were assigned manually to the 1971 to 1977 segment of the file and computer assigned to the 1978 to mid-1982 segment. The paper is divided into three parts: section one provides background information; section two considers the method used to develop the system and its design; and section three discusses the application of the codes to existing ABI/INFORM citations. (Author/DMC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

David K. Bond

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED245703

IR050 1776

DEVELOPMENT AND IMPLEMENTATION OF A HIERARCHICAL CLASSIFICATION SYSTEM FOR THE ABI/INFORM DATABASE

David K. Bond, Data Courier Inc.

ABSTRACT

The ABI/INFORM database includes abstracts from more than 500 journals covering all aspects of business and management. It is published by Data Courier Inc. and includes material dating back to 1971. Because of the diverse nature of the database and the imprecision of management terminology, it was determined that online searchers would be well-served by a hierarchical classification system. Such a system has been developed and implemented over the past two years.

Because of the costs involved, it is common in the online industry for such enhancements to be searchable only from the date of implementation. The management of Data Courier Inc., however, believed this feature should operate throughout the database. Classification codes were assigned manually to the 1971 to 1977 segment of the file and computer assigned to the 1978 to mid-1982 segment. This paper will discuss the overall development and implementation of the classification system.

INTRODUCTION

Since online bibliographic databases first became practical they have undergone almost constant change. Rapid advances in computer technology have dramatically increased the speed at which these databases can be searched and information retrieved. These "hardware" advances have been accompanied by expanded search software capabilities. Today, online database vendors such as DIALOG Information Services, System Development Corporation (SDC) and Bibliographic Retrieval Services (BRS) present users with an almost endless number of access points for literally hundreds of databases. These databases range from scientific and technical files such as Engineering Index and Chemical Abstracts to more general files in the social sciences and business.

Armed with these very powerful capabilities, online searchers expect to get complete and accurate information quickly and easily. This presents publishers of bibliographic databases with the challenge of designing features that take advantage of search systems' capabilities to enhance online information retrieval.

This paper discusses the development and implementation of a search feature to meet this challenge. The feature is a hierarchical classification system for the ABI/INFORM business database. Section one provides background information. Section two considers the method used to develop the system and its design. Section three discusses how the codes were applied to existing ABI/INFORM citations.

BACKGROUND

ABI/INFORM is a business and management database which abstracts from more than 500 business-related journals. As of April 1983, the file contains over 200,000 records. Journals included in the file cover diverse management disciplines and industries and range from specialized academic journals to broad-interest general-business magazines. To further complicate matters, users of ABI/INFORM have interests as varied as the material included in the file. One user might want to retrieve everything written on quality circles while another might want to search only accounting publications for the implications of a specific piece of legislation. To better accommodate our users and the material we abstract and index, we needed some means of creating various subsets of the database. A hierarchical classification system addresses this need.

Another factor contributing to the need for a classification system is the imprecise nature of business and management terminology. This imprecision allows a word to have many different meanings depending on the context in which it is used. "Options," for example, can be defined as alternative methods or purchase contracts, and promotion has very different connotations to advertising and personnel managers. This imprecision is not a particular problem in traditional print indexes where subheadings can be used for clarification. However, in computer searching a database of over 200,000 citations, this imprecision can create considerable problems. A classification system that would allow online searchers to create subject-specific subsets would eliminate much of this imprecision.

A final factor contributing to the need for a classification system concerns the indexing philosophy of the ABI/INFORM editors. We developed a comprehensive (approximately 8,000 terms) thesaurus and decided to index at the most specific level. We do not always assign very broad terms such as ACCOUNTING and DATA PROCESSING to all articles that deal with these subjects. We feel that this style of indexing is less cluttered and remains the most appropriate style for online information retrieval. However, it does create problems for online searchers who want to subdivide the database into various subject categories. For example, to create a subset of all articles in the database on data processing, one must connect many related terms with the Boolean operator OR. With a numerical classification system, such segmenting of the database is easily accomplished by entry of a four-digit code.

DEVELOPMENT OF THE CLASSIFICATION SYSTEM

The ABI/INFORM Editorial department began by looking at various classification schemes for business literature. We found schemes used by other business databases of little use. The most useful references were the listing of subject headings used by the American Management Association in its Index to AMA Resources of the Seventies 1970-1976

compiled by Elizabeth A. Keegan (AMACON, 1977), and The London Classification of Business Studies, Second edition by K. D. C. Vernon and Valerie Lang (Aslib, 1979).

As a result of this initial work, a preliminary list of categories and subcategories was developed. These codes were assigned, on a test basis, to articles entering the database. During the test period, the classification system was revised periodically. Developing the classification codes by applying them to actual articles entering the database proved to be a good approach. We were able to spot areas that required more detail and to formulate guidelines for code assignment at the same time. The ABI/INFORM indexing staff participated in developing the classification system. These same people were later responsible for assigning classification codes to new records entering the database.

After spending approximately one year developing the codes, we were satisfied that they worked very well for the material we would be classifying. We were, however, not certain that the overall system would be perceived as clear and logical by online searchers. We retained Warner-Eddison Associates, Inc., Cambridge, Massachusetts, to critique the classification system. They compared our classification system with others and recommended changes to make it more useful to online searchers. Warner-Eddison's report was received in April, 1982 and the classification system became final in June of that year.

DESIGN OF THE CLASSIFICATION SYSTEM

To encompass the breadth of the ABI/INFORM database we identified five distinct types of codes.

- Three define business environment, management functions, and industries and markets. These codes identify subject content.

- Two identify article treatment and organization types. These identify the type of material being classified and the types of organizations under discussion.

The classification codes are four digits, thus the system allows up to four levels of specificity. Appendix 1 contains the complete classification system.

The ABI/INFORM database contains a wealth of information on such areas as economic conditions, international trade, energy resources and economic, social and energy policy. This information is very important to business decision-making but is not directly linked to any one management function. The business environment codes make it easy to retrieve this kind of information.

Management function codes identify areas common to all businesses. Often they can be related directly to departments within corporations. A major focus of the ABI/INFORM database is on these functional areas. Searchers can use the management function codes to create complete subsets of any subject area in the file. For example, a searcher looking for trends in employee benefits can choose the 6400 category (employee benefits & compensation) to retrieve everything in the database on employee benefits. Searches can then be refined further by date limiting or with index terms.

Searchers of ABI/INFORM often need to limit search results to a specific industry. For example, a client may want to know how a specific management technique has been applied in the chemical industry. Industry and market codes identify approximately 30 agricultural, extractive, service and manufacturing industries. Codes for the financial services and insurance industries are assigned at two levels while other industry codes are not. Since these two industries are extensively covered in ABI/INFORM, more space is devoted to them in the classification system. Industry codes are assigned only to articles that have an industry focus.

Treatment codes define the type of article being classified. Assignment of these codes is straightforward and their use can be accurately discerned from the code names. One very useful treatment code is "non-US." It is assigned to articles that discuss companies, industries, laws, economic conditions, etc. in countries other than the US. This code will help searchers most by allowing them to exclude such information from search results.

Organization codes are assigned to articles that focus on a particular type of organization. For example, the code for public sector organizations can be used to separate articles dealing with public administration from those discussing corporate and nonprofit management.

The ABI/INFORM classification system was designed for online information retrieval. Therefore, all codes can be combined using the Boolean operators AND, OR and NOT. For example, a searcher interested in retrieving articles on legislation affecting the retail banking industry can specify the code 8120 (retail banking) AND the code 4320 (legislation). To broaden searches, the codes can be truncated. If the searcher in the above example wanted to broaden the search to include the entire financial services industry, that could be easily accomplished by truncating after the 81. The results would then include commercial banking, retail banking or investment services.

ASSIGNING CLASSIFICATION CODES TO EXISTING ABI/INFORM CITATIONS

Once the classification codes were finalized, we had to decide whether to make the codes a searchable feature throughout the

database or only from the date we started assigning them.

We started assigning classification codes in July 1982. At that time the ABI/INFORM database contained approximately 170,000 records. Of these 170,000, the approximately 65,000 that entered the database from 1971 to 1977 had been indexed by a computer program. This auto-indexing did not meet current standards, and we had already committed to manually reindexing that segment of the database. The staff of temporary indexers hired to perform that task were able to assign classification codes as well.

The 105,000 records that entered the database from 1978 to mid-1982 presented a problem. Manual assignment of classification codes to each of these records was not financially feasible. However, since these citations had been indexed manually, automatic assignment of classification codes based on index terms and source journals seemed a workable approach.

In analyzing the possibilities of "autoclassifying," we found that, in many cases, we were unable to get the required accuracy at the third and fourth levels of specificity. We decided to assign business environment and management function codes to the second level, and industry and market codes to the fourth level. Only one treatment code (company specific) and one organization code (public sector) were assigned by the autoclassification process. For consistency, indexers used these same levels in classifying the 1971 to 1977 records.

In specifying the criteria for automatic assignment of classification codes, the editorial department developed exhaustive index-term search strategies for each code. Creating these strategies involved a tradeoff between the complexity of the autoclassifying software we would have to write and the accuracy we wanted to achieve. Assignment of a code (or codes) was based on the presence of one of the following:

- A specified single index term
- Two specified index terms
- A specified journal code.

With these conditions in mind, we formulated initial strategies. We looked first for journals from which all articles could be classified with specified codes. Some economic theory, management science and certain organizational behavior journals fall into this category. We then analyzed each term in the ABI/INFORM thesaurus and assigned classification codes to all subject-specific terms. For example, the term ACCOUNTING PROCEDURES was assigned the code for accounting; the term BANK MARKETING was assigned the code for the banking industry and the code for marketing. We entered these journal codes and index terms along with their classification codes into our word processor, sorting them numerically by classification code. This gave a preliminary strategy for each code. (See fig. 1)

Figure 1

Excerpt from the strategy used to automatically assign the classification code 1100 (Economics). The three letter codes identify source journals.

1100	*JXB
1100	*RES
1100	*RSU
1100	*SEJ
1100	ACCELERATION PRINCIPLE (ECON)
1100	ADMINISTERED PRICES (ECON)
1100	AGRICULTURAL ECONOMICS
1100	ANTI INFLATION
1100	ANTI INFLATION ACT-CANADA
1100	ANTITRUST
1100	ANTITRUST LAWS
1100	APPROPRIATE TECHNOLOGY
1100	AVERAGE PROPENSITY TO CONSUME (ECON)
1100	AVERAGE PROPENSITY TO SAVE (ECON)
1100	BALANCE OF PAYMENTS
1100	BALANCE OF PAYMENTS DEFICIT
1100	BALANCE OF TRADE
1100	BERNOULLI HYPOTHESIS (ECON)
1100	BILATERAL AID
1100	BLACK MARKETS (COM)
1100	BRETTON WOODS SYSTEM
1100	BUSINESS CYCLES
1100	CANADIAN DOLLAR
1100	CAPITAL INTENSIVE INDUSTRIES

We tested each strategy on the DIALOG system by using the Boolean operator OR to link index terms and journal codes. We then limited these searches to citations from 1978 to mid-1982 and printed a sample of about ten percent. For each sample, we printed the index terms and article titles obtaining a random sample by entering several different type commands.

We analyzed each sample to isolate

- Other terms that should have been included in the strategy
- Terms that had to be deleted from the strategy because they produced false drops
- Pairs of terms that, when assigned individually, were not always specific enough to warrant a class code, but when assigned together accurately indicated a code.

Each autoclassifying strategy was reviewed, tested and revised several times before final strategies were developed.

After all strategies were finalized and corrections had been made, we ran the autoclassifying program. Approximately ninety-seven percent of the records were assigned at least one code. An average of 2.15 codes was assigned to each record. (Manual classification of the 1971 to 1977 records resulted in a similar number of codes per record.) Classification codes were assigned manually to records which were not picked up by the autoclassifying strategies.

We attained a high degree of accuracy, but because we used an automated process, some inaccuracy was inevitable. For some codes, such as those covering Human Resource Management, assignment was nearly perfect. For other less specific codes, such as Investment Analysis, assignment was somewhat less accurate.

Developing the autoclassifying strategies involved a tradeoff familiar to all online searchers--recall versus relevance. In most cases, we achieved high levels of accuracy in both areas. In cases where, because of the nature of the classification code, we had to sacrifice recall or relevance, we structured the strategies for maximum recall. We felt that recall was more important because classification codes represent only one access point to the database. As long as complete recall is assured, searchers can refine search strategies with other features, such as index terms or journal names.

CONCLUSION

Online searchers have a powerful new feature with which to retrieve information from the ABI/INFORM database. The codes allow searchers to easily create exhaustive subject-specific subsets of the database. This will help achieve higher recall and relevance. Searchers will be able to select, for example, the code for advertising and be confident that the set of articles retrieved contains everything in the database that deals with advertising; thus higher recall. They can then use the Boolean operator AND to combine this set with other classification codes, index terms or words in the abstract or title and be certain that they will retrieve only articles about advertising; thus, higher relevance.

APPENDIX 1

ABI/INFORM CLASSIFICATION SYSTEM

BUSINESS ENVIRONMENT

1100 Economics
- *1110 Economic conditions & forecasts
- *1120 Economic policy & planning
- *1130 Economic theory

1200 Social policy

1300 International trade & foreign investment (see also multinational corporations 9510)
- *1310 Foreign investment in the US

1500 Energy/environment
- *1510 Energy resources
- *1520 Energy policy
- *1530 Natural resources
- *1540 Pollution control

MANAGEMENT FUNCTIONS

2000 GENERAL MANAGEMENT

2100 Administrative & management personnel
- *2110 Board
- *2120 CEO
- *2130 Executives

2200 Managerial skills

2300 Planning & strategy
- *2310 Planning
- *2320 Organizational structure (subsidiaries, decentralization, etc.)
- *2330 Acquisitions & mergers

2400 Public relations
- *2410 Social responsibility
- *2420 Image
- *2430 Business-government relations

2500 Organizational behavior

2600 Management science/operations research

3000 FINANCE

3100 Capital & debt management

3200 Credit management

3300 Risk management

3400 Investment analysis

3500 Foreign exchange administration (see also 1120)

3600 Pension fund management

4000 ACCOUNTING, TAXATION, & LAW

4100 Accounting
- *4110 Accounting firms & accountants
- *4120 Accounting policies & procedures
- *4130 Auditing

4200 Taxation
- *4210 Institutional taxation
- *4220 Estate planning
- *4230 Personal taxation

4300 Law
- *4310 Regulation
- *4320 Legislation
- *4330 Litigation

5000 OPERATIONS

5100 Facilities management
- *5110 Office management
- *5120 Purchasing
- *5130 Maintenance
- *5140 Security
- *5150 Energy management
- *5160 Transportation

5200 Communications & information management
- *5210 Office automation
- *5220 Data processing management
- *5230 Hardware
- *5240 Software & systems
- *5250 Telecommunications systems
- *5260 Records management

5300 Production management
- *5310 Production planning & control
- *5320 Quality control
- *5330 Inventory management
- *5340 Safety management

5400 Research & development

6000 HUMAN RESOURCE MANAGEMENT (see also 2500)

6100 Human resource planning

6200 Training & development

6300 Labor relations

6400 Employee benefits & compensation

6500 Employee problems

7000 MARKETING

7100 Market research

7200 Advertising

7300 Sales & selling

7400 Distribution

7500 Product planning & development (see also research & development 5400)

(Appendix 1 continued)

INDUSTRIES & MARKETS

8100 Financial services industry
- 8110 Commercial banking
- 8120 Retail banking
- 8130 Investment services

8200 Insurance industry
- 8210 Life & health
- 8220 Property casualty

8300 Other services
- 8310 Consultants (not elsewhere classified)
- 8320 Health care
- 8330 Broadcasting & telecommunications (not equipment)
- 8340 Electric, water, & gas utilities
- 8350 Transportation (not equipment)
- 8360 Real estate
- 8370 Construction & related
- 8380 Hotel & restaurant
- 8390 Retail stores, includes groceries

8400 Agriculture

8500 Extractive
- 8510 Petroleum industry

8600 Manufacturing
- 8610 Food processing, includes beverages & liquors
- 8620 Textile & apparel industries
- 8630 Lumber & wood products, includes paper
- 8640 Chemicals, includes rubber & plastics
  - 8641 Pharmaceuticals
- *8650 Electrical, electronics, instrumentation
  - 8651 Computers
- 8660 Metalworking
- 8670 Machinery (industrial, construction, farm, etc.)
- 8680 Transportation equipment (cars, aircraft, ship-building, etc.)
- 8690 Publishing

ARTICLE TREATMENT

- 9110 Company specific/case studies
- *9120 Product specific
- *9130 Experimental/theoretical
- *9140 Statistical data
- *9150 Guidelines
- *9160 Biographical
- *9170 Non-US
- *9180 International

ORGANIZATION TYPES

- *9510 Multinationals: companies with subsidiaries and/or operations in more than one country
- *9520 Small business
  - 9521 Minority- and women-owned business
- *9530 Diversified companies
- *9540 Non-profit institutions
- 9550 Public sector

BEST COPY AVAILABLE